Ecology, 86(12), 2005, pp. 3398–3410

REARRANGEMENT OF ECOLOGICAL DATA MATRICES VIA MARKOV CHAIN MONTE CARLO SIMULATION

ISTVÁN MIKLÓS, IMELDA SOMODI, AND JÁNOS PODANI[1]

Department of Plant Taxonomy and Ecology, Eötvös Loránd University, Pázmány Péter sétány 1/c, H-1117 Budapest, Hungary

Abstract. Block clustering and seriation reveal the underlying structure of ecological data structures by rearranging the rows and the columns of the data table or data matrix, usually representing species and sample sites, respectively. Classical approaches to this problem rely upon a goodness criterion optimized through an iterative algorithm or utilize a simultaneous classification or ordination of species and sites. The new procedure introduced here does not strive for a single optimal rearrangement. Instead, it generates a series of probability distributions, the Boltzmann distributions, for a large set of solutions that include both the optimal and suboptimal solutions. The procedure is governed by a hypothetical parameter, *T,* called the ''temperature.'' For the value of zero, only the best rearrangements (if there are many) have nonzero probabilities. As the value of this parameter increases, less optimal rearrangements become more frequent in the associated series of distributions. When *T* approaches infinity, the distribution becomes uniform over all the possible matrix rearrangements. We propose a Markov chain Monte Carlo (MCMC) method which converges reasonably fast to this distribution for any value of *T.* This chain provides a sample of matrices that can be characterized via several statistics based on the Boltzmann distribution.

The relevance of the method is demonstrated using ecological data. We illustrate how much extra information can be gained from suboptimal solutions that may have biological meaning not revealed by the best solution. Although the objective is to give a distribution of solutions rather than a single optimal solution, the new method can actually outperform heuristic searching algorithms in finding the best arrangement.

We provide source code for the MCMC method in the C language, which can be compiled under many operating systems (Windows, Linux/Unix, Macintosh OS) and used in command line mode. The Linux/Unix version operates in interactive mode, it gives a graphical output of the results, and is available as a web server interface in PHP 4.3, which can also be installed on personal computers.

Key words: *aquatic vegetation; block clustering; Boltzmann distribution; MCMC; seriation.*

INTRODUCTION

Multivariate data analysis has long been an integral part of numerical ecology, with advantages and merits treated in considerable detail in the literature (Digby and Kempton 1987, Legendre and Legendre 1999). Classification and ordination procedures provide results in the form of mathematical constructs such as dendrograms or other forms of tree graphs and scatter diagrams including biplots and triplots (Lepš and Šmilauer 2003). In addition, there are procedures that operate directly on the raw data matrices by rearranging their rows and columns (usually representing species and sites, respectively) to emphasize and optimally visualize underlying data structures (Podani 2000). The rearranged data may then speak for themselves and may represent valuable alternatives to mathematical objects, such as trees and ordinations. Fig. 1 illustrates the major possibilities of this approach. Block clustering methods explore the data matrix for the presence of distinguishable and homogeneous data blocks, whereas seriation places emphasis on mutually optimal orderings of rows and columns. A hybrid procedure, called block seriation, places data blocks along the diagonal of the matrix to facilitate reciprocal gradient interpretability of species and site groups. Interpretation of the rearranged data matrix is usually enhanced by shading, that is, the original values are categorized and these categories are depicted in different shades from white to black (McIntosh 1978). Algorithms for matrix rearrangement utilize a statistical criterion that may be either global (for example, sum of squares or chi square, see Feoli and Orlóci [1978], Podani and Feoli [1991]) or based on some local properties such as the dissimilarities/distances between neighboring rows or columns. Optimization of these measures of goodness of data block homogeneity (block sharpness) or of seriation represents the core of matrix rearrangement algorithms.

Finding the optimal solution of matrix rearrangement is not a simple task. Many problems in this field are

Manuscript received 7 January 2005; revised 27 April 2005; accepted 3 May 2005. Corresponding Editor: N. G. Yoccoz.

[1] Corresponding author. E-mail: podani@ludens.elte.hu

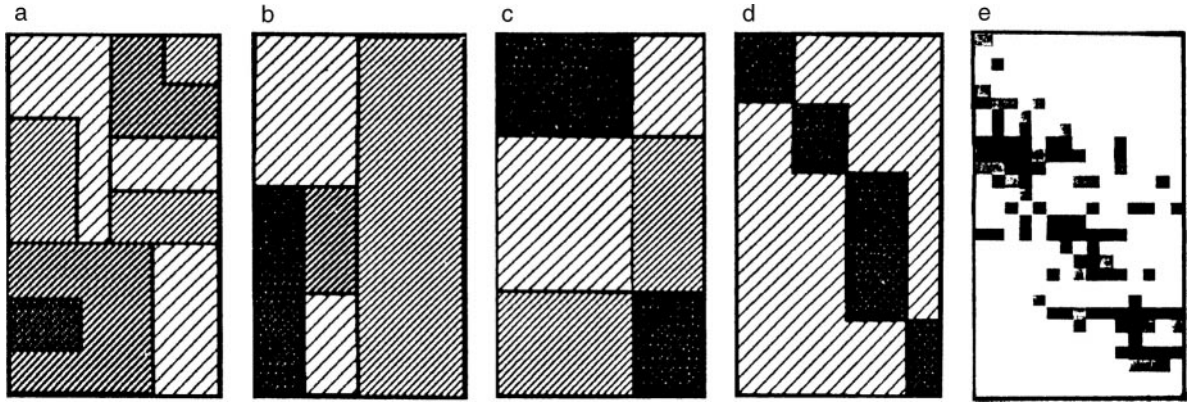

FIG. 1. Main types of matrix rearrangement and shading problems: (a) unconstrained and (b) partial block clustering, (c) cross partitioning, (d) block seriation, and (e) seriation of raw data. Shading reflects within-block homogeneity (a–d) or deviation from 0 (e). The figure is modified from Podani (2000), with permission from Backhuys Publishers, Leiden, The Netherlands.

NP-complete, an epithet referring to the mathematical fact that there is no algorithm to find the optimum in a reasonable time for practical problem sizes. Heuristic greedy approximations can be used instead, but one can never be sure that the optimum has been found, even after thousands of runs of the searching algorithm. Even if an optimal solution is obtained, it may happen that the optimum is not unique and the alternatives remain undetected. Furthermore, the optimum-oriented approach deliberately neglects suboptimal solutions that might be just as informative biologically as the optimal ones. We feel that examination of the distribution of results produced by many runs of the greedy algorithm offers a more flexible approach to the problem and is more exhaustive than searching for absolute optima. This was first suggested by Podani and Feoli (1991), who examined the frequency distribution of global and local optima in the case of iterative block clustering procedures via cross partitions (Fig. 1c). However, such a distribution of results is greatly influenced by the greedy strategy and is not suitable for statistical characterization because different greedy strategies yield practically incomparable distributions.

In this paper, we suggest a new approach to investigate data structuring. We define a continuous series of Boltzmann distributions of matrix rearrangements. These distributions have a hypothetical ''temperature'' parameter. At zero temperature, the distribution is uniform for the optimal matrix rearrangements (if there is more than one). With increasing temperatures, less optimal solutions have larger and larger probabilities of occurring and, eventually, on infinitely high temperatures, the distribution becomes uniform over all the possible matrix rearrangements. We sample from these distributions through a Markov chain Monte Carlo technique (MCMC; Liu 2001). In this, a very long sequence of matrices is generated by randomly modifying the matrix from one step to the other. The general theorem of MCMC guarantees that the chain is always forced to converge to the prescribed distribution irrespective of the nature of random modifications. This is an essential difference between MCMC and naïve randomization approaches. In particular, a matrix is retained with a given probability in the series even if the random modification deteriorates the goodness criterion, whereas greedy strategies are forced to improve the result as much as possible in every step.

This paper is concerned with the unconstrained block clustering and seriation approaches (Fig. 1a and e). For unconstrained block clustering, we define the optimality criterion based on the absolute differences between neighboring rows and columns. Furthermore, between-column and between-row similarity indices are proposed for characterizing the distribution sampled. An ecological example is used to demonstrate that several optimal solutions may exist simultaneously, and that statistics based on a sample from optimal solutions (statistics on zero temperature) are more informative than a single optimal solution. Also, statistics based on higher temperatures, i.e., those derived from samples involving suboptimal solutions, reveal useful information that cannot be obtained from optimal solutions. We demonstrate the utility of plexus graphs in summarizing and visualizing the between-column and between-row relationships in the actual set of solutions.

To show the general applicability of our approach, we also considered the seriation problem. In this case, the optimality criterion is defined based on the Robinson property (Robinson 1951). Here, the Boltzmann distribution of matrix rearrangements is informative of the probability of a species or a site to occur in a given row or column, respectively, which cannot be displayed via plexus graphs. Therefore, we introduce another efficient visualization technique, the distance distribution of rows and columns. We also show that the new method outperforms greedy algorithms described earlier (Podani 1994). It is in accordance with the well-known fact that MCMC-based stochastic search, known as

simulated annealing, is usually superior to greedy optimization (Aarts and Korst 1989).

METHODS

Energy of a matrix

Various optimality criteria have been suggested to achieve different matrix rearrangement objectives (Hartigan 1975, Podani 2000, and references therein). A measure of the goodness of rearrangement is called here the energy function, denoted by E_π and Ψ_π for matrix rearrangement π in case of unconstrained block clustering and seriation, respectively. The probability distribution of matrix rearrangements is defined via their energy (see *The Boltzmann distribution of matrix rearrangements*). In this distribution, ''better'' configurations have greater probability and ''worse'' rearrangements have lesser probability. Any appropriate energy function must have the following properties. First, energy is inversely proportional to the goodness of rearrangement. This ensures that the better a given solution, the greater the probability of the matrix (Kirkpatrick et al. 1983). In addition, we propose that the energy function should be scale independent. This ensures that multiplying every entry in the matrix by a scalar does not change the distribution.

In this paper, we use two energy functions, one for unconstrained block clustering and the other for seriation, as described in the following sections.

A criterion for unconstrained block clustering.—Here, the objective of matrix rearrangement is to concentrate large data values into any number of blocks of any size and shape, thus visualizing the mutual relationships of variables and objects in terms of maximally homogeneous clusters of numbers. For this purpose, we suggest using the sum of the sum of absolute differences between neighboring rows and between neighboring columns as the optimality criterion. The fact that this measure considers only local properties of the matrix makes it suitable for concentrating large values into unconstrained blocks. As a starting point in deriving a goodness measure, let us consider the sum of absolute differences (Manhattan distances) between neighbors, given by the formula

$$E_\pi = \sum_{1\le i\le m}\sum_{1\le j\le n-1}\left|x_{\pi(i,j)} - x_{\pi(i,j+1)}\right| + \sum_{1\le i\le m-1}\sum_{1\le j\le n}\left|x_{\pi(i,j)} - x_{\pi(i+1,j)}\right| \quad (1)$$

where $\pi(i, j)$ denotes the ith row and jth column in the rearranged matrix. This measure, however, suffers from the imbalance that rows and columns on the border of the matrix receive less emphasis than those within the matrix, a phenomenon called the border effect.

Border effect.—Each entry inside the data matrix has four neighbors, the entries at the border have three, and those at the corners have only two. Since we define the energy of a matrix in terms of Manhattan distances from four neighbors, entries at the border require specific treatment to compensate for the fewer number of comparisons. In addition to the case without any correction (a), we can choose from three solutions (b–d) for correcting the border effect and thus facilitating four comparisons for all positions.

a) Disregard borders. We do only three or two comparisons for border entries as implied by Eq. 1, thus forgetting about the entire problem. This method tends to place high values to the border of the matrix, especially if it is hard to find similar neighbors for these values.

b) Zero-frame. The matrix is surrounded by a frame of 0 values. These zeros are fixed, i.e., no rearrangement can remove them from these positions. Since border values are compared with zeros, this solution tends to move small values to the border.

c) Torus. The first and the last rows are treated as neighbors, just like the first and the last columns. This way we indirectly assume the presence of a cyclic gradient in both ways, which is usually not the case.

d) Mirror. We define a frame as being equal to the second and the penultimate rows and columns. This method implies double weighting of the distance between the boundary vectors and their neighbors which seems reasonable, because these are the only proposed neighbors of the bordering values.

The first three procedures are burdened by consistent bias towards a particular type of result, as demonstrated by the artificial examples of Fig. 2 as well. To the contrary, in the mirror technique the different rearrangements will have different frames thus removing the bias present in the previous procedures. Therefore, the energy function (Eq. 1) is modified as follows:

$$E_\pi = \sum_{1\le i\le m}\sum_{1\le j\le n-1}\left|x_{\pi(i,j)} - x_{\pi(i,j+1)}\right| + \sum_{1\le i\le m-1}\sum_{1\le j\le n}\left|x_{\pi(i,j)} - x_{\pi(i+1,j)}\right| + \sum_{1\le j\le n}\left[\left|x_{\pi(1,j)} - x_{\pi(2,j)}\right| + \left|x_{\pi(m-1,j)} - x_{\pi(m,j)}\right|\right] + \sum_{1\le i\le m}\left[\left|x_{\pi(i,1)} - x_{\pi(i,2)}\right| + \left|x_{\pi(i,n-1)} - x_{\pi(i,n)}\right|\right]. \quad (2)$$

The above criterion can be used in the optimization of rearrangements of a given matrix. However, this is affected by the scale on which the actual data are measured and energy values obtained for different matrices cannot be compared. To resolve this problem, E_π can be divided by the average absolute differences between neighbors,

$$\mathrm{av}_X = \frac{\sum_{1\le i\le m}\sum_{1\le k<l\le n}\left|x_{i,k} - x_{i,l}\right| + \sum_{1\le k\le n}\sum_{1\le i<j\le m}\left|x_{i,k} - x_{j,k}\right|}{m\binom{n}{2} + n\binom{m}{2}} \quad (3)$$

where $x_{i,k}$ is an entry of the $m \times n$ matrix $\mathbf{X}$ being analyzed. In this way, we have derived a measure which

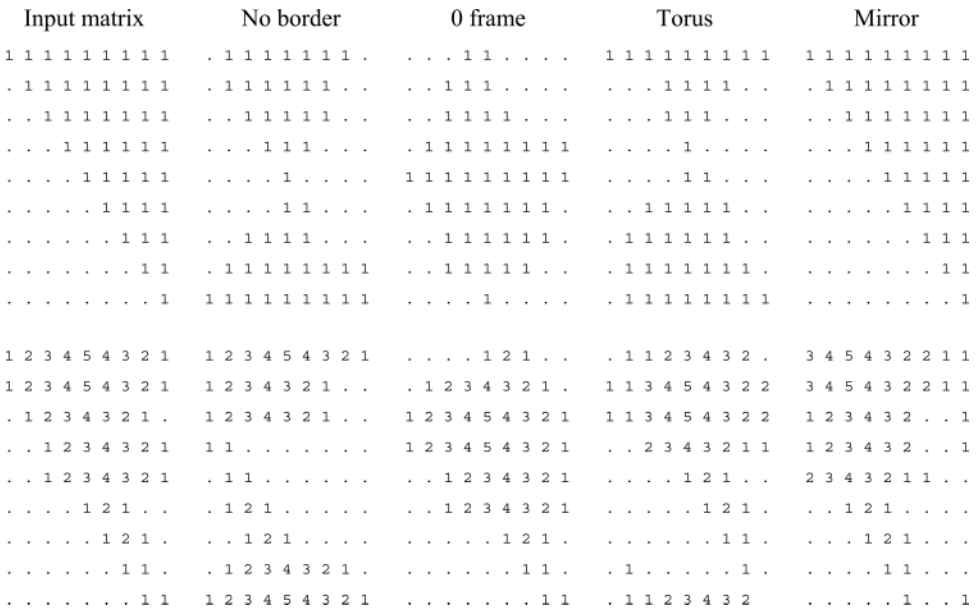

FIG. 2. Two small artificial data matrices (left) and the most optimal solutions obtained by four different methods of border effect correction.

satisfies the two fundamental requirements for an energy function.

A seriation criterion.—The classical seriation problem (Kendall 1970, 1971) is to find a simultaneous ordering (or permutation) of the rows and the columns of the data matrix with the objective of revealing a background one-dimensional gradient. The basic idea is that large scores should be concentrated around the main diagonal as closely as possible, whereas low values should fall as far from it as possible. This goal is best achieved by considering the so-called Robinson property (Robinson 1951). A matrix is said to have this property if its values decrease monotonically in the rows and the columns when moving away from the main diagonal in both directions. To express deviation from this perfect state numerically, Podani (1994) suggested the following energy function:

$$\Psi_\pi = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{\pi(ij)}\left[\left|\frac{n\times i}{m} - j\right| + \left|\frac{m\times j}{n} - i\right|\right]. \quad (4)$$

In this formula, each data entry is weighted by the sum of its positional differences from the diagonal, i.e., the number of rows and columns through which the value has to be moved to reach the diagonal. This weight is not an integer in most situations. An advantage of Eq. 4 is that no border effect is present. To achieve scale independence, Ψ_π can also be divided by the sum of entries in the matrix, so that both requirements are met.

The Boltzmann distribution of matrix rearrangements

The classical optimality requirement for matrix rearrangement is to minimize the energy value. In Appendix A, we show that this problem is NP-complete at least for the unconstrained case, therefore there is no fast algorithm which always finds the best solution unless P = NP (that is, unless a universal algorithm exists which quickly solves a very large set of computationally hard problems). No such universal algorithm has been constructed so far. Mathematicians tend to believe that such an algorithm does not exist at all, although no proof has been given to confirm this view.

Instead of finding a single optimal solution, we define the so-called Boltzmann distribution (see for example, Liu 2001) of matrix rearrangements based on the energy function E or Ψ as

$$P_T(\pi) \propto e^{-E_\pi/T} \quad \text{or} \quad P_T(\pi) \propto e^{-\Psi_\pi/T} \quad (5)$$

where P_T denotes the probability density function (pdf) for temperature T and $\propto$ means ''proportional to.'' We cannot calculate analytically the normalizing constant for P_T; and we shall show later that it is not necessary anyway for sampling from $P_T(\pi)$. The temperature T introduced here does not have any biological meaning, but we can describe its qualitative effect. In the Boltzmann distribution, the probability of worse rearrangements decreases exponentially, and on temperature $T = 1/\ln(k)$, a rearrangement whose energy is one unit larger than the optimal rearrangement is k times less probable than the optimal rearrangement. When $T \to 0$ or, equivalently, when $k \to \infty$, only the minimum energy rearrangements have nonzero probability, and the distribution is the uniform one on these rearrangements (if there are more than one). As $T \to \infty$ (hence $k \to 1$), the distribution converges to the uniform distribution of all the possible matrix rearrangements. Experience suggests that a little above zero temperature,

suboptimal solutions dominate the distribution; namely, the distribution ''melts'' quickly.

Efficient characterization and visualization of results

There are $m!n!$ different rearrangements for matrix $\mathbf{X}_{m,n}$. (If we consider a permutation equal to its reverse, then the number of possible matrices reduces to $n!m!/4$, but this does not influence the subsequent discussion.) Although most of these rearrangements have a negligible probability on small temperatures, the number of good candidates may be still too large. Some subsequent analysis is needed to facilitate joint evaluation of these alternative results which efficiently compresses the information carried by them. It is a natural attempt to characterize the distribution of matrices based on the optimization criterion. Therefore, we develop statistics capturing the goodness of neighborhoods of columns and rows for the block clustering solutions, as well as statistics estimating the probability of a species or a site to occur at a given distance from the middle of the matrix for the seriation results.

Similarity of neighboring rows or columns: creating a plexus graph.—We introduce indices of similarity between pairs of rows and between pairs of columns. These measures will be useful in summarizing unconstrained block clustering results in form of a plexus graph. Let $r_\pi(i, j)$ denote the row neighbor-indicator function, which is 1 if rows i and j of the original matrix are neighbors in the matrix rearrangement π, otherwise it is 0. Similarly, $c_\pi(k, l) = 1$ if columns k and l of the data matrix are neighbors in the matrix rearrangement π, otherwise this function yields 0. The similarity of rows i and j at temperature T is defined as the probability that these two rows are neighbors in the distribution P_T, and is given by the following function:

$$\mathrm{Sr}_T(i, j) = \sum_\pi P_T(\pi) r_\pi(i, j). \tag{6}$$

In an analogous manner, the similarity of columns k and l at temperature T is defined as

$$\mathrm{Sc}_T(k, l) = \sum_\pi P_T(\pi) c_\pi(k, l). \tag{7}$$

Note that the values of these indices range from 0 to 1, since they are probabilities. A value of 0 means that the two rows or columns in question are never neighbors in matrices that have nonzero probability in the distribution, while 1 means that they are always neighbors.

We can characterize a distribution at a given temperature with two plexus graphs with vertices representing species in the first one and sites in the second one. Two vertices are connected if the similarity between the corresponding species or sites exceeds a given threshold. A more elaborate approach involves use of edges of different thickness to indicate the similarity level (cf. McIntosh 1978). To facilitate easy comparison, we draw the vertices along a circle and fix their order as one of the permutations obtained in a matrix rearrangement with the smallest energy value. This choice is arbitrary if there are several optimal solutions.

Characterizing the Boltzmann distribution for the seriation criterion.—Let us now define the center of a data matrix $\mathbf{X}_{m,n}$ as being the position $m/2,n/2$ which may be actually a value for odd values of both m and n, or a position between values otherwise. It is easy to see that the energy function given by Eq. 4 is invariant for mirroring the matrix on its center. Therefore, the relevant information is how far a row or column is from the center of the matrix in a given rearrangement. Let $\mathrm{rd}_\pi(i, d)$ denote the row-distance indicator function, which is 1 if row i is d positions away from the center of the matrix in rearrangement π, otherwise $\mathrm{rd}_\pi(i, d) = 0$. Similarly, let $\mathrm{cd}_\pi(i, d)$ denote the column-distance indicator function. We can define the distance probability distribution of row i on temperature T:

$$\mathrm{Prd}_T(i, d) = \sum_\pi P_T(\pi)\mathrm{rd}_\pi(i, d) \tag{8}$$

where $\mathrm{Prd}_T(i, d)$ is the probability that row i is d positions far from the center of the matrix in the Boltzmann distribution on temperature T. Similarly we can define the distance probability distribution of column j on temperature T:

$$\mathrm{Pcd}_T(j, d) = \sum_\pi P_T(\pi)\mathrm{cd}_\pi(j, d). \tag{9}$$

These distributions for all rows or all columns can be plotted in a three-dimensional diagram. The row or column indices of the original matrix are shown on the x-axis, in an order based on an optimal rearrangement. Due to the mirror invariance discussed above, pairs of rows or columns being from the same distance from the center of the matrix in the optimally rearranged matrix are plotted next to each other, namely, the first and last rows or columns are neighbors, etc. The y-axis is for the distances measured from the center of the matrix, and we measure the probabilities $\mathrm{Prd}_T(i, d)$ or $\mathrm{Pcd}_T(j, d)$ on the z-axis. If this optimal rearrangement had probability 0.5 in the Boltzmann distribution (hence its mirror had the other 0.5), we could see a straight diagonal distribution on this chart. Deviations from the diagonal reveal the possibility of alternative rearrangements in the Boltzmann distribution.

The Markov chain Monte Carlo method

The only remaining question is how to obtain the Boltzmann distribution. Since $m!n!$ is an exceedingly high number even for moderate problem sizes, brute force calculation is not feasible. Therefore, we suggest estimating the distribution and the derived similarity indices through sampling from the distribution P_T using a Markov chain Monte Carlo (MCMC) method. In MCMC, it is unnecessary to determine the probability density function exactly (Liu 2001), which would only be possible if the normalizing constant were known.

Therefore, the formula in Eq. 5 is indeed sufficient for the purpose.

The simulations start from a random permutation of rows and columns. In each step, a part of the permutation for the actual rearrangement π_{old} is perturbed to create a proposal, π_{new}. When the Boltzmann distribution is defined using the unconstrained block clustering criterion, a random part of the permutation is inverted. The starting and the last elements of this inversion are determined randomly, and the decision whether this is done with the rows or the columns is also random. Inversion is illustrated by the example below:

$$1\ 4\ 3\ 6\ 5\ 8\ 2\ 9\ 7 \rightarrow 1\ 4\ 8\ 5\ 6\ 3\ 2\ 9\ 7$$

in which we inverted the section [3 6 5 8]. It has been shown (e.g., Cameron 1999) that, for permutations longer than three, such operations provide an ergodic chain on all the possible permutations, a fundamental requirement for Markov chain Monte Carlo methods (Liu 2001). For the seriation criterion, π_{new} is obtained by swapping two randomly chosen, not necessarily neighboring columns or rows. These moves define an irreducible but periodic Markov chain (only an even number of moves can cancel each others' effect out [Cameron 1999]). This means that the primary chain thus defined is not ergodic. However, the probability for rejecting a move is non-zero. Since rejection breaks the periodicity, the final Markov chain will be ergodic.

To determine whether the newly proposed matrix is retained in the chain we use the Metropolis ratio (Metropolis et al. 1953, Liu 2001). For the energy function E, this takes the following form

$$\frac{\exp\left(\frac{-E_{\pi_{new}}}{T}\right)}{\exp\left(\frac{-E_{\pi_{old}}}{T}\right)} = e^{-\Delta E/T} \tag{10}$$

where $\Delta E = E_{\pi_{new}} - E_{\pi_{old}}$. For seriation, E is replaced by Ψ in the above formula. Then, a uniform random number on (0, 1) is generated and π_{new} is retained if this number is smaller than the above ratio. Otherwise, π_{new} is rejected and the new member of the Markov chain will be π_{old}. This random choice on accepting or rejecting π_{new} guarantees convergence to the predefined distribution (Metropolis et al. 1953, Liu 2001). For sampling from the Boltzmann distribution, we define three arbitrary integers, B, I, and S. To eliminate autocorrelation, we consider only every Ith matrix in the sequence, whereas the number of such matrices is defined as S. At the beginning, the first B matrices are discarded to ensure convergence to the desired distribution (known as the burn-in phase in the MCMC literature). Therefore, the total length of the Markov chain is $B + I \times S$, yielding a sample of S matrices.

The time complexity of matrix perturbation and the associated computations is relatively low. For block clustering, an inversion in the row permutation changes only $O(n)$ neighbor relationships, whereas an inversion in the column permutation changes only $O(m)$ neighbor relationships. We do not need to rewrite matrix entries themselves in the Markov chain, only the corresponding permutations, therefore a sampling step requires only $O(n + m)$ computational time. In the case of seriation, swapping two columns or two rows and calculating the energy difference between the original and the newly obtained matrix can also be performed in $O(n + m)$ time.

We would like to emphasize that the type of perturbation does not affect the distribution to which the Markov chain converges, only the speed of convergence and mixing time are influenced. The reason for introducing different moves for the two different criteria is that mixing by inversions was found to be very slow for the seriation criterion, while mixing was much faster when matrices were perturbed with swapping columns and rows. Nevertheless, calculating the energy difference for inversion-based mixing in seriation would take $O(nm)$ time. That is, computational complexity would be significantly greater than for mixing via swapping.

Obviously, some of the proposals will be significantly worse than the actual rearrangement, and they will be accepted only with very small probability. Therefore, a straightforward alternative strategy would calculate the energy of all possible proposals available from the actual rearrangement, and would propose the best solutions with the highest probability. Such an approach could be seen as the discrete version of Monte Carlo updating method based on Langevin diffusion (Roberts and Rosenthal 1998). This technique has been successfully applied in other MCMC methods of seriation (Buck and Sahu 2000). However, our empirical results showed that our plain Metropolis-Hastings algorithm has a better overall performance, due to avoiding the high computational cost of inferring all possible proposals in each step of the Markov chain.

DATA

The data set used to demonstrate the advantages of using a distribution of solutions originates from vegetation samples collected for the comparison of three floating island complexes, called sites. Two old floating fens, developed by primary succession, and young floating islands developed by secondary succession were included. All the sites are located in Hungary. One of the old floating fen complexes is in Lake Velencei (47°10′ N, 18°32′ E), the other one in several adjoining oxbows of the smaller branch of the Danube south of Budapest (48°45′ N, 19° E). The young floating islands developed following artificial flooding of a former primary floating fen complex drained ca. 50 years ago. The study site was located in the center of a secondary floating island formation (Ingó) within the shal-

TABLE 1. Vegetation strata used in sampling.

Stratum no.	Name	Dominant species	Discriminating features
1	peat moss–willow scrub	*Sphagnum fimbriatum*, *Salix cinerea*	presence of *Sphagnum fallax*, *Sphagnum squarrosum*, *Thelypteris palustris*
2	willow scrubs and lesser reedmace beds	*Salix cinerea* and/or *Typha angustifolia* and *Thelypteris palustris*	absence of *Sphagnum* species
3	reed stands	*Phragmites australis*	presence of *Solanum dulcamara*, *Eupatorium cannabinum*
4	large sedge beds	*Carex pseudocyperus*, *Carex elata*, *Carex riparia*	presence of tall forb species (*Bidens cernuus* and *Rumex maritimus*)
5	forb vegetation on muddy surfaces	*Bidens cernuus*, *Rumex maritimus*, *Cyperus fuscus*	presence of *Chenopodium ficifolium*, *Epilobium hirsutum*, *Veronica anagallis-aquatica*

low lake of Kis-Balaton (47°30′ N, 17°10′ E), situated southwest of Lake Balaton.

At each site, the sampling design was stratified (Podani 2000, Sheldon et al. 2002) over the dominant vegetation groups present in the floating fens to cover the full variation and most of the potentially available gradients. Several vegetation types distinguished by dominant species or species groups were present in the area. We used five strata (Table 1) which were based on these species dominance patterns. Within each stratum, a minimum of five 25-m^2 quadrats were placed to estimate percent cover of species (vascular plants, bryophytes, and charophytes; Appendix C). To sum up, 31 quadrats were placed at Ingó, 41 in Lake Velencei, and 29 in the Danube oxbows. The full data matrix contains 78 species and 101 quadrats. For further details on the study sites, strata, and data collection methods, consult Somodi and Botta-Dukát (2004).

RESULTS

Unconstrained block clustering

Monte Carlo sequences of matrix rearrangements were generated at 20 different temperatures, from $T = 0.001$ to $T = 0.041$. When the energy function after $B = 10\,000$ steps decreased to the fluctuating stage (''burn-in''), we retained $S = 10\,000$ rearrangements, and to diminish autocorrelation, there were $I = 10\,000$ steps in the Markov chain between two matrices that were retained. This very exhaustive investigation could be performed in less than 12 h on an Intel Pentium 4 2.0 GHz computer under a RedHat 8.0 operating system. Here, we describe detailed results for two temperatures. We chose the lowest investigated temperature, $T = 0.001$, where most of the rearrangements have the minimum energy. At the other temperature chosen, $T = 0.017$, suboptimal solutions already dominate the Boltzmann distribution. At higher temperatures, only the strongest connections remain in the plexus graph, indicating that the Boltzmann distribution approaches a uniform distribution (see supplementary information in Appendix B containing a .gif movie of plexus graphs for a temperature range from $T = 0.001$ to $T = 0.041$).

At temperature $T = 0.001$, 9502 of the sampled matrices had the smallest energy value (25 006). All these smallest energy matrices were different. This confirms good mixing of the Markov chain, and also shows that there are numerous optimal rearrangements in this case. We chose one of the optimal rearrangements (Fig. 3) to define the order of species and quadrats in the plexus graphs. The threshold of similarity (for Eqs. 6 and 7) to be shown in the plexus graphs was set to be the value exceeding 10 times the average similarity.

Most of the edges of the plexus graphs run around the circle, reflecting that many neighbor relationships correspond to a one-dimensional structure. However, there are several edges connecting objects (species or quadrats) which are not neighbors, and there are several neighbors that are not connected. The former indicates that in other optimal or suboptimal solutions these objects can be neighbors quite frequently, and the latter means that although these objects are neighbors in this particular solution, this neighborhood is only one possibility among other solutions. These observations are in accordance with our biological knowledge, as discussed below.

Quadrats originating from the same sampling strata, which were identified formerly as being representatives of coherent vegetation types (see Somodi and Botta-Dukát 2004), are mainly grouped along the circle. The only exception is stratum 2, which is subdivided into three groups according to minor dominance differences (Fig. 4). Two of these subgroups can clearly be associated with dense *Salix cinerea* scrubs, while the third is characterized by lower willow cover and higher cover of *Thelypteris palustris*. Reed stands with relatively high proportion of *Typha angustifolia* were also divided from the main body of reed bed vegetation. These finer distinctions were undetected by correspondence analysis and were therefore disregarded when broad vegetation types were identified (Somodi and Botta-Dukát 2004).

Seven species groups of various size can be recognized along the circle of the optimal arrangements of species (Fig. 5). Most of these groups can be associated

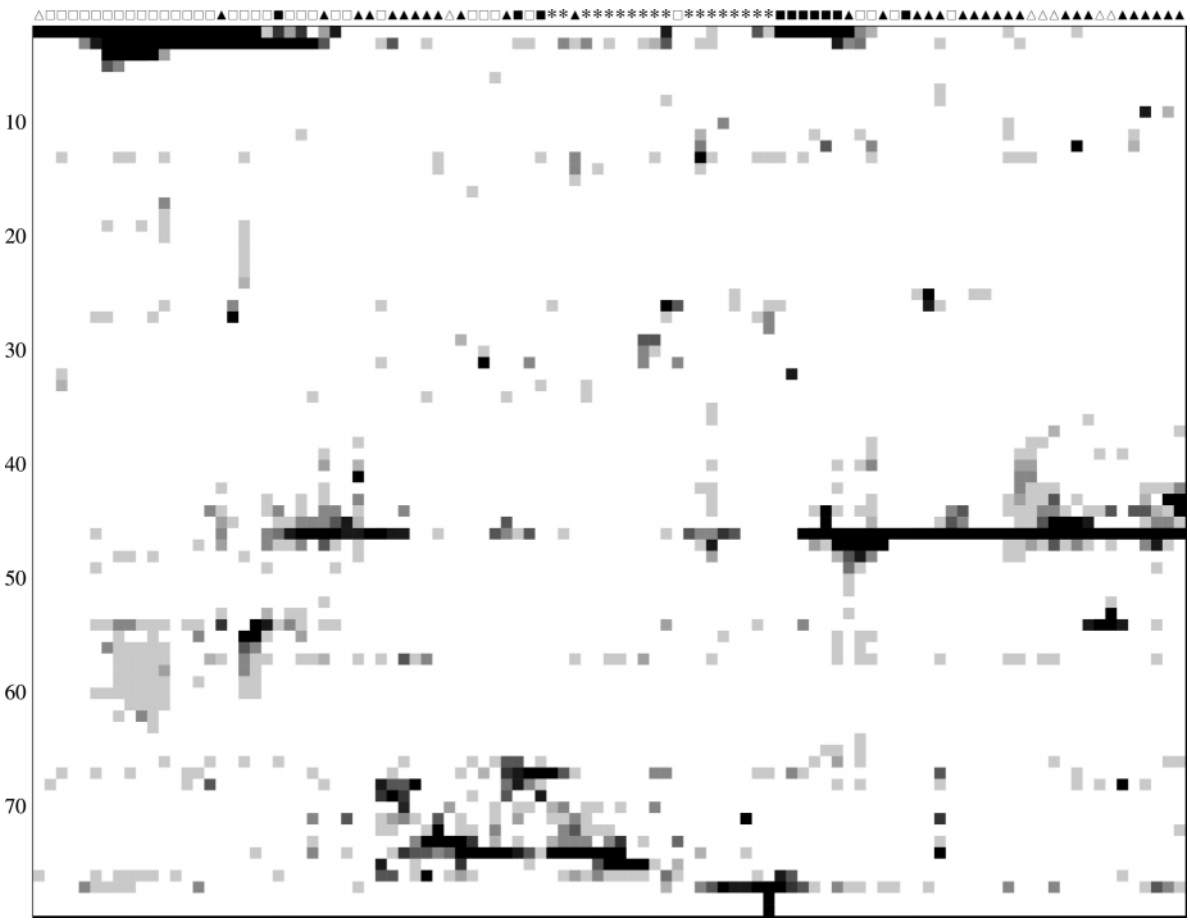

FIG. 3. An optimal rearrangement of aquatic vegetation data based on the unconstrained block-clustering criterion. The *y*-axis shows species numbers as they appear in Appendix D. Symbols for sites: open triangles, peat moss–willow scrub; solid triangles, willow scrubs and lesser reedmace beds; open squares, reed stands; solid squares, large sedge beds; asterisks, forb vegetation on muddy surfaces. Shading is proportional to percent cover, from white (0%) to black (100%).

with a stratum, others with a site. This also helps to differentiate between species characteristic of a certain vegetation type and those characteristic of sites. The peat moss–willow scrub, the tall forb community, the dwarf mud vegetation (the last two are subtypes of the vegetation type belonging to Stratum 5), and the large sedge beds (Table 1) appear to have characteristic species combinations. The Danube oxbows have a characteristic set of species of their own, while reed stands at Lake Velencei also seem to have a separate species group. Floating species behave similarly, though they can be associated with neither sites nor strata. It is important to note that neither of these combinations is based on the dominant species themselves. These combinations can be considered characteristic of the vegetation groups mentioned.

There are edges connecting species that are not neighbors along the circle at the lowest temperature as well. For example, in the optimal arrangement, *Sphagnum squarrosum* belongs to the species characteristic of the Danube oxbows, and is also linked to a group characteristic of peat moss–willow scrubs.

When temperature is increased to $T = 0.017$, the plexus graph disintegrates because many similarity values fall below the threshold (Figs. 4b and 5b). At the same time, at this temperature new edges appear that did not show up at lower temperatures. A potential reason behind this phenomenon is that there might be several suboptimal rearrangements containing the same pattern of neighbors. At low temperatures, the cumulative probabilities for the high number of suboptimal rearrangements are still smaller than the probability of the optimal rearrangement(s). However, on higher temperature, the difference between the probability of an optimal and a suboptimal rearrangement is smaller, so that the sum of probabilities of suboptimal rearrangements may exceed the total for the optimal ones.

We can explain the appearance of new edges on biological grounds. In both the quadrat and the species graphs, the new edges that appear on higher temperatures carry additional information. In the case of quadrats, for example, some of them from large sedge beds are separated from the main group of large sedge bed quadrats. If the optimal solution were examined only,

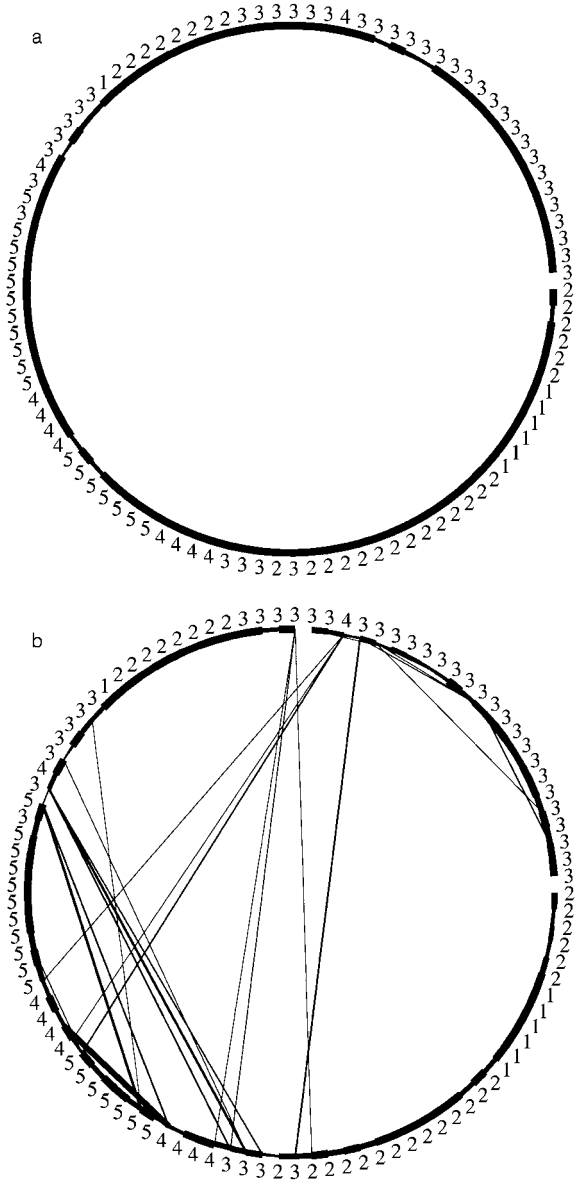

FIG. 4. Plexus graphs of the quadrats at temperatures (a) $T = 0.001$ and (b) $T = 0.017$. Quadrats are arranged based on the optimal solution depicted in Fig. 3. See Table 1 for numbering of quadrats.

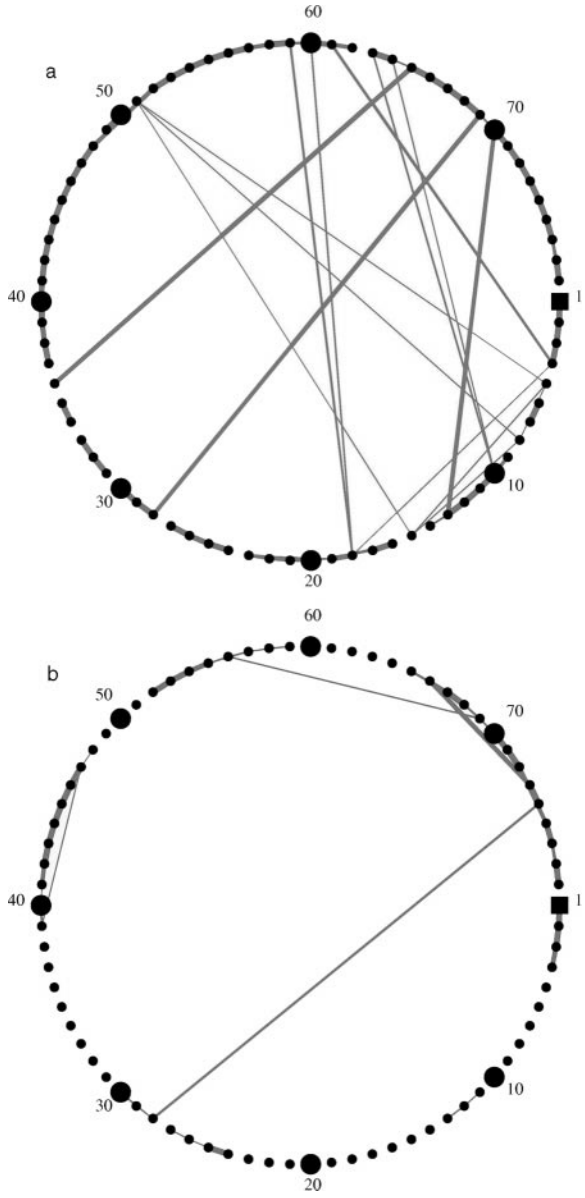

FIG. 5. Plexus graph of the species at temperatures (a) $T = 0.001$ and (b) $T = 0.017$. Species are arranged based on the optimal solution depicted in Fig. 3. For clarity, every 10th species is emphasized by large symbols. For plant names, see Appendix C.

this separation would leave the similarity of these quadrats undetected. Nevertheless, the new edges on higher temperatures connect these few quadrats to the large group of sedge-dominated quadrats, which confirms the integrity of this vegetation type. In terms of species, the new edge between the floating species (*Lemna minor* and *Utricularia vulgaris*) and dwarf forbs of muddy habitats (e.g., *Veronica anagallis-aquatica*) is due to a special co-occurrence pattern. The periodical intrusion of water onto the mud surfaces of the young floating islands sparsely inhabited by dwarf species carries along the floating species, mainly *Lemna minor*. These floating plants can survive on the wet mud, thus causing the apparent association observed. The fact that this edge appears only on higher temperatures shows that this association receives relatively low support. Another example is the edge between *Lindernia procumbens* and *Cyperus fuscus*. In the best arrangement, *L. procumbens* is part of the species group characteristic of dwarf mud vegetation, while *C. fuscus* primarily belongs to the species group corresponding to the tall forb vegetation. *C. fuscus* is present in both vegetation subtypes, in the first one as a dominant species, in the

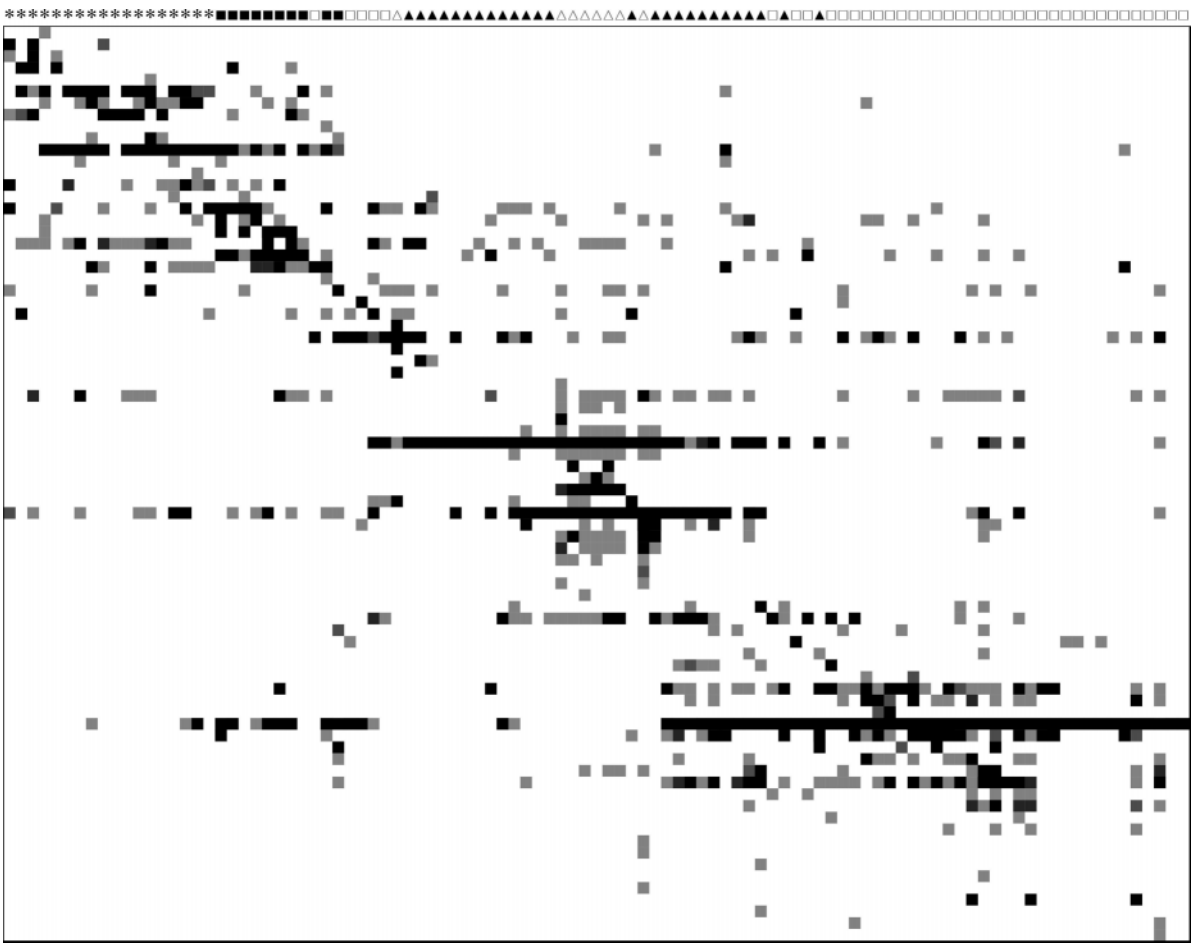

FIG. 6. One of the most optimal solutions for seriating the aquatic vegetation data set. Symbols for sites are the same as in Fig. 3. Species numbering is different from that in Fig. 5 and is not shown to avoid confusion. Shading is proportional to percent cover, from white (0%) to black (100%).

second one as an ''understory'' beneath the tall forbs. The edge represents the strong association of the species with the dwarf mud vegetation.

Seriation

We examined the Boltzmann distribution with the seriation criterion on two temperatures, $T = 0.001$ and $T = 0.017$. We used 10 000 steps for the burn-in phase, and then every 10 000th rearrangement was retained resulting in a total of 10 000 matrices. At temperature $T = 0.001$, 7974 of the rearrangements had the minimum energy value ($\Psi = 161\,762.01$, one result shown in Fig. 6) followed by 1857 rearrangements having a slightly greater energy ($\Psi = 161\,762.19$). As many as 9990 matrices had energy smaller than $\Psi = 161\,763$, and among them 9971 were different. This shows good mixing and indicates the possibility of large number of optimal solutions. Remarkably, an iterative greedy algorithm (Podani 1994) performs worse than MCMC. In three days of computational time, the analyses converged into 100 different solutions, and the best rearrangement had an energy of $\Psi = 167\,537.4$. We must mention though that the comparison of running times of the new and old methods is not fair since the old method was implemented in FORTRAN, and the program was run under a WINDOWS XP operating system (Microsoft, Redmond, Washington, USA) on a 3-GHz processor. However, the huge differences between computational times (3 d vs. <1 h) and between the performance of the two methods deserve appreciation.

At temperature $T = 0.017$, the smallest energy value obtained was $\Psi = 161\,765$. However, the distribution still shows the main properties of the best rearrangement (see Fig. 7 for the species). Most of the species have only a limited possibility to be rearranged and only the rare species have uncertain positions. We found even lower variability for quadrats (diagram not shown).

The biological interpretation of results is very similar to that of block clustering. Species groups characteristic of strata are distinguishable here as well, though other species groups like floating species or the

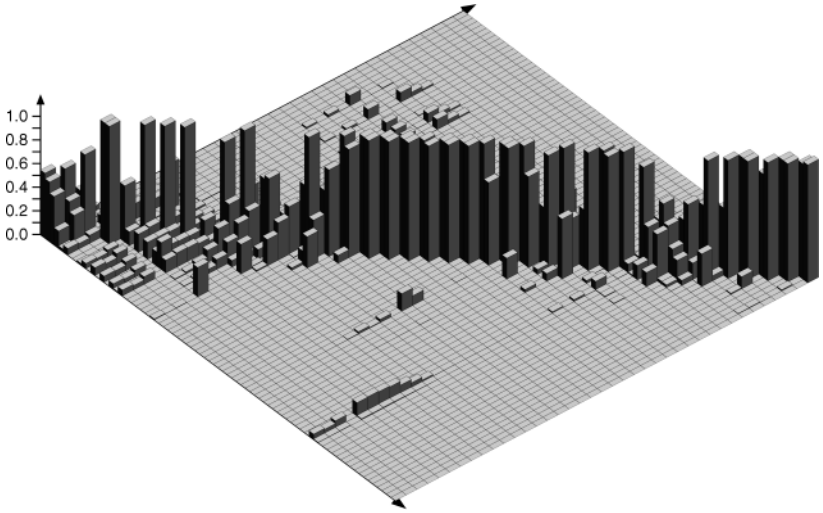

FIG. 7. The Prd(i, d) distribution (the probability that row i is d positions from the center of the matrix in the Boltzmann distribution) at temperature $T = 0.017$ for species (rows).

group assigned to one of the sites are lacking. At the bottom of the rearranged matrix (Fig. 6), a mixed cluster appeared in which presumably uninformative species are collected. An advantage of block clustering against seriation is that uncorrelated species are less influential. Quadrats from each stratum also clumped closely together; this clumping is even stronger than what we observed in the best rearrangement after block clustering. For example, two subgroups of quadrats from large sedge beds are only joined in block clustering after suboptimal rearrangements had been taken into account, while in the seriation result all of the quadrats from large sedge beds appeared in the same group. An advantage of the seriation approach is that it provides a plausible mutual correspondence between species and quadrat groups. Therefore, seriation could serve as a first approach to finding broad correspondence between quadrat and species groups, while block clustering results including suboptimal solutions could provide deeper insight into the strength and pattern of associations. The gradient implied by seriation has to be handled with care, though. The transitions along the gradient have no obvious biological relevance in this example, except for overlaps between the groups.

DISCUSSION

We introduced a new approach to the rearrangement of ecological data matrices. Instead of searching for a single optimal solution, we defined a series of distributions: the Boltzmann distributions of matrix rearrangements. In this distribution, the probability of a matrix rearrangement decreases exponentially with the deviation of the energy (measure of goodness) from the energy of the best rearrangement. We showed that this approach is applicable to unconstrained block clustering and seriation if energy is defined by our rearrangement criteria, and there is no doubt that it should just as well work for other measures as well. Though it is hard to handle these distributions analytically, we showed that the MCMC approach allows efficient sampling from these distributions. For both rearrangement problems, we also introduced statistics efficiently characterizing and visualizing the Boltzmann distributions: plexus graphs for unconstrained block clustering and the distance distributions of rows and columns for seriation.

We tested our method on actual ecological data, and showed that:

1) There are potentially several rearrangements that are optimal according to a given criterion. A single optimal solution is less informative than a distribution of many solutions in which the optimal solutions have the greatest probabilities.

2) Suboptimal solutions are also useful and reveal information not conveyed by optimal solutions. Suboptimal solutions enable, for example, the recognition of overlapping groups of species or quadrats and also generate a temperature-driven ''melting'' hierarchy of objects. Unlike in many types of classification and tabular rearrangement results, in plexus graphs no object is forced into a single group, similar to overlapping clustering or clumping. An example is *Sphagnum squarrosum*, which belongs to a species group associated with one of the three sites, but also belongs to one of the vegetation types.

A further advantage of the proposed method with respect to biological interpretation is its ability to reveal species or their groups which are characteristic combinations in the sample. Species weakly associated with other species can be regarded as less informative. In this way, potential characteristic species can be identified and the number of species involved in further analyses can be effectively reduced. However, finding and interpreting relationships between species groups

and quadrat groups (vegetation types) require expert knowledge.

Although our objective was to get a distribution of rearrangements rather than a single best solution, the latter can be highlighted from the sampled rearrangements. We showed that our approach can actually outperform heuristic techniques both in computational time needed to get a good solution and in the goodness of the best solution found. We would like to mention the high similarity between our approach and a stochastic optimization procedure called simulated annealing (SA; Kirkpatrick et al. 1983). SA is also based on the Boltzmann distribution but temperature decreases while the Markov chain proceeds, thus freezing the chain into the best solution. An advanced MCMC technique inspired by SA is parallel tempering (PT; Geyer 1991, Hukushima and Nemoto 1996). In PT, several Markov chains run simultaneously, converging to Boltzmann distributions of the same type but on different temperatures. Stochastic communication between the chains provides faster mixing for all chains without significant increases of computational time, while keeping the convergence for each chain. Since we suggest investigating the Boltzmann distributions on several temperatures, PT is the definite choice for this purpose in the future.

We also showed that finding the best rearrangement for unconstrained block clustering is NP-complete, and we conjecture similar results for other matrix rearrangement problems. Hence, we can never be sure that the best solution has been found by any approach, including ours. However, a set of suboptimal solutions also reveals the main properties of the best solution. Indeed, the Boltzmann distributions on higher temperatures almost never contain any optimal solution, however, the introduced statistics on higher temperatures were comparable with those on lower temperature.

We did not define any prior distribution of temperatures and did not use any prior information on how a good rearrangement should look like. With such priors, we would be able to put our work into a Bayesian modeling framework. Such attempts have already been made in archaeology, for example, by introducing prior assumptions on the measurement error (Buck and Sahu 2000) or on dynamics of cultural changes (Halekoh and Vach 2004). Another potentially interesting approach is the minimum message length method (Wallace and Boulton 1968, Dale and Dale 2004), where the objective is to minimize the sum of the description length of the prior information and the likelihood. Finding possible ways to incorporate prior knowledge into our model is a promising challenge.

Markov chain Monte Carlo methods have been the basic statistical tools in many fields in biology (Liu 2001, Larget 2004), and their power was also already shown for some statistical problems in ecology (Zaman and Simberloff 2002, Miklós and Podani 2004). The main drawbacks of MCMC are that it requires expert knowledge in statistics and that general purpose MCMC software packages are still lacking. With this paper, we wanted to facilitate the spread of MCMC approaches in ecological data analysis and to provide a software package for the techniques introduced here (Supplement).

Acknowledgments

István Miklós is supported by a Békésy György postdoctoral fellowship. He also wishes to thank Jotun Hein, Alexei Drummond, and Gerton Lunter for learning about MCMC. János Podani was supported by the Hungarian Research Fund (OTKA) grant no. SUP043732. We would like to thank the anonymous referees for their valuable comments that, we feel, greatly improved the paper.

Literature Cited

Aarts, E. H. L., and J. Korst. 1989. Simulated annealing and Boltzmann machines: a stochastic approach to combinatorial optimisation and neural computing. Wiley, Chichester, UK.

Buck, C. E., and S. K. Sahu. 2000. Bayesian models for relative archeological chronology building. Journal of the Royal Statistical Society D **49**:423–440.

Cameron, P. 1999. Permutation groups. Cambridge University Press, New York, New York, USA.

Dale, M. B., and P. E. R. Dale. 2004. Sources of uncertainty in ecological modelling: predicting vegetation types from environmental attributes. Community Ecology **5**:203–225.

Digby, P. N. G., and R. A. Kempton. 1987. Multivariate analysis of ecological communities. Chapman and Hall, London, UK.

Feoli, E., and L. Orlóci. 1978. Analysis of concentration and detection of underlying factors in structured tables. Vegetatio **40**:49–54.

Geyer, C. J. 1991. Markov chain Monte Carlo maximum likelihood. Pages 156–163 *in* E. Keramigas, editor. Computing science and statistics: the 23rd symposium on the interface. Interface Foundation, Fairfax, Virginia, USA.

Halekoh, U., and W. Vach. 2004. A Bayesian approach to seriation problems in archaeology. Computational Statistics and Data Analysis **45**:651–673.

Hartigan, J. A. 1975. Clustering algorithms. Wiley, New York, New York, USA.

Hukushima, K., and K. Nemoto. 1996. Exchange Monte Carlo method and application to spin glass simulations. Journal of the Physical Society of Japan **65**:1604–1608.

Kendall, D. G. 1970. A mathematical approach to seriation. Philosophical Transactions of the Royal Society of London, Series A **269**:125–135.

Kendall, D. G. 1971. Seriation from abundance matrices. Pages 215–252 *in* F. R. Hodson, D. G. Kendall, and P. Tautu, editors. Mathematics in the archaeological and historical sciences. Edinburgh University Press, Edinburgh, UK.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. Science **220**:671–680.

Larget, B. 2004. Introduction to Markov chain Monte Carlo methods in molecular evolution. Pages 49–66 *in* R. Nielse, editor. Statistical methods in molecular evolution. Springer series in statistics for biology and health. Springer-Verlag, New York, New York, USA.

Legendre, P., and L. Legendre. 1999. Numerical ecology. Second edition. Elsevier, Amsterdam, The Netherlands.

Lepš, J., and P. Šmilauer. 2003. Multivariate analysis of ecological data using CANOCO. Cambridge University Press, Cambridge, UK.

Liu, J. S. 2001. Monte Carlo strategies in scientific computing. Springer series in statistics. Springer-Verlag, New York, New York, USA.

McIntosh, R. P. 1978. Matrix and plexus techniques. Pages 159–191 *in* R. H. Whittaker, editor. Ordination and classification of communities. Junk, The Hague, The Netherlands.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. Journal of Chemical Physics **21**:1087–1091.

Miklós, I., and J. Podani. 2004. Randomization of presence/absence matrices: comments and new algorithms. Ecology **85**:86–92.

Podani, J. 1994. Multivariate data analysis in ecology and systematics. SPB Academic Publishing, The Hague, The Netherlands.

Podani, J. 2000. Introduction to the exploration of multivariate biological data. Backhuys, Leiden, The Netherlands.

Podani, J., and E. Feoli. 1991. A general strategy for the simultaneous classification of variables and objects in ecological data tables. Journal of Vegetation Science **2**:435–444.

Roberts, G. O., and J. S. Rosenthal. 1998. Optimal scaling of discrete approximations to Langevin diffusions. Journal of the Royal Statistical Society B **60**:255–268.

Robinson, W. S. 1951. A method for chronologically ordering archaeological deposits. American Antiquity **16**:131–147.

Sheldon, F., A. J. Boulton, and J. T. Puckridge. 2002. Conservation value of variable connectivity: aquatic invertebrate assemblage of channel and floodplain habitats of a central Australian arid-zone river, Cooper Creek. Biological Conservation **103**:13–31.

Somodi, I., and Z. Botta-Dukát. 2004. Determinants of floating island vegetation and succession in a recently flooded shallow lake, Kis-Balaton (Hungary). Aquatic Botany **79**: 357–366.

Wallace, C. S., and D. M. Boulton. 1968. An information measure for classification. Computer Journal **11**:185–194.

Zaman, A., and D. Simberloff. 2002. Random binary matrices in biogeographical ecology—instituting a good neighbor policy. Environmental and Ecological Statistics **9**:405–421.

APPENDIX A

A proof demonstrating that the energy problem of matrix rearrangement is NP-complete is available in ESA's Electronic Data Archive: *Ecological Archives* E086-187-A1.

APPENDIX B

A .gif movie showing the change of plexus graphs when temperature is changed is available in ESA's Electronic Data Archive: *Ecological Archives* E086-187-A2.

APPENDIX C

A list of species is available in ESA's Electronic Data Archive: *Ecological Archives* E086-187-A3.

SUPPLEMENT

A software package for performing tabular rearrangement via MCMC is available online in ESA's Electronic Data Archive: *Ecological Archives* E086-187-S1.